MMDB: Entrez's 3D structure database

Aron Marchler-Bauer, Kenneth J. Addess, Colombe Chappey, Lewis Geer, Thomas Madej, Yo Matsuo, Yanli Wang and Stephen H. Bryant*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received November 3, 1998; Accepted November 5, 1998

ABSTRACT

The three dimensional structures for representatives of nearly half of all protein families are now available in public databases. Thus, no matter which protein one investigates, it is increasingly likely that the 3D structure of a homolog will be known and may reveal unsuspected structure–function relationships. The goal of Entrez's 3D-structure database is to make this information accessible and usable by molecular biologists (http://www.ncbi.nlm.nih.gov/Entrez). To this end Entrez provides two major analysis tools, a search engine based on sequence and structure 'neighboring' and an integrated visualization system for sequence and structure alignments. From a protein's sequence 'neighbors' one may rapidly identify other members of a protein family, including those where 3D structure is known. By comparing aligned sequences and/or structures in detail, using the visualization system, one may identify conserved features and perhaps infer functional properties. Here we describe how these analysis tools may be used to investigate the structure and function of newly discovered proteins, using the PTEN gene product as an example.

AN EXAMPLE: THE PTEN GENE PRODUCT

Germline mutations in the PTEN gene product have been shown to cause Cowden Disease, and somatic mutations have been associated with a variety of cancers (1). To illustrate the analysis tools of Entrez's (2,3) 3D-structure database we describe how it may be used to answer the following question: is the 3D structure known for the PTEN gene product, or a homologous protein, and does this information suggest mechanisms by which these mutations might cause disease? To use this article as a tutorial readers should address their WWW browser to the Entrez site (http://www.ncbi.nlm.nih.gov/Entrez) and perform the analysis step by step, as we describe it.

USING SEQUENCE NEIGHBORS TO FIND 3D-STRUCTURES

To retrieve the sequence of the PTEN gene product we enter Entrez's Pubmed literature database and type the query 'PTEN and Cowden Disease'. This identifies a number of articles describing mutations in the PTEN gene and their associated phenotypes. Choosing 'Display Protein Links' leads one to sequences reported in these articles, and in particular to the Swiss-Prot (4) entry PTEN_HUMAN. Studying the associated GenPept Report we see annotations recording many of the mutations reported in the literature. In particular, we see that mutations at residues 123, 124, and 129 in the PTEN gene product have been implicated in Cowden Disease.

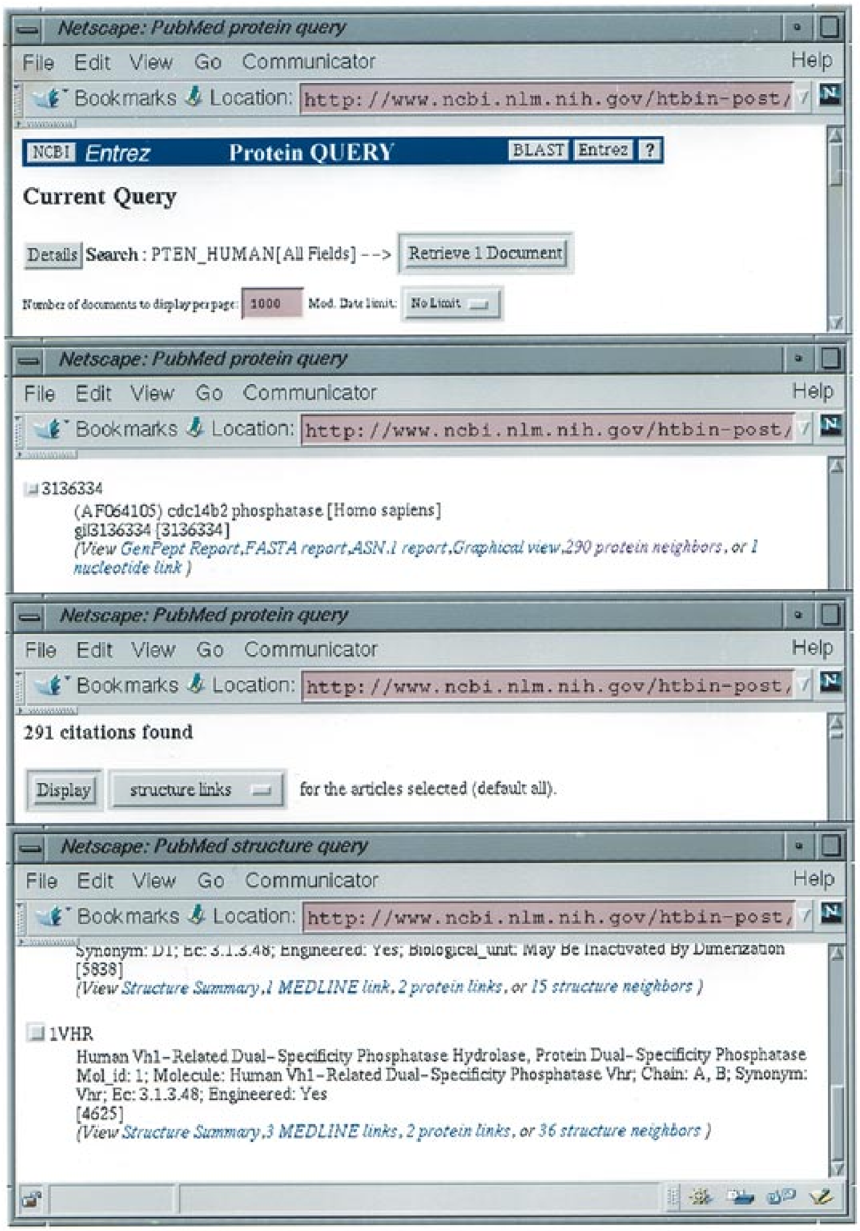
The sequence PTEN_HUMAN does not have a Structure Link, since Entrez has collected this entry from Swiss-Prot, not PDB (5). To find a structure we need to search among sequences similar to PTEN, and to do so we choose its Protein Neighbors. This link retrieves all sequences with significant similarity to PTEN_HUMAN, the results of the pre-computed BLAST (6) searches that comprise Entrez's sequence neighbor database. To see if the 3D structure is known for any of these homologous sequences one may browse this list, looking for a Structure Link, or choose to Display Structure Links for all of these sequences. Again the results are negative, indicating that none of the sequences detected as similar to PTEN in a single round of BLAST neighboring has 3D structure.

At this point one might conclude that no 3D structure relevant to PTEN is known, and indeed one has learned that there is no close homolog with 3D structure. One may use Entrez to continue the search with greater sensitivity, however, by examining the 'neighbors of neighbors' of PTEN. A strategy to follow is to browse the list of PTEN's sequence neighbors, searching for a sequence that is annotated as having the same function as PTEN, and at the same time a larger number of sequence neighbors. Following this strategy one skips over PTEN neighbors that are large multi-domain proteins, such as kinases containing tensin-like domains. One identifies Cdc14b2, however, a human protein 202 residues in length, which, like PTEN, is annotated as a phosphatase. Cdc14b2 has nearly 300 sequence neighbors, and Display Structure Links yields a hit: 1VHR, the 3D structure of human VH1-related dual-specificity phosphatase (7). This search strategy is illustrated in Figure 1.

VIEWING STRUCTURES AND STRUCTURE NEIGHBOR ALIGNMENTS

A useful first step in identifying structure–function relationships suggested by a structure is to examine it by molecular graphics.

*To whom correspondence should be addressed. Tel: +1 301 435 7792; Fax: +1 301 480 9241; Email: bryant@ncbi.nlm.nih.gov


Netscape: PubMed protein query
File Edit View Go Communicator Help
Bookmarks Location: http://www.ncbi.nlm.nih.gov/htbin-post/
NCBI Entrez Protein QUERY BLAST Entrez ?

Current Query

Details Search : PTEN_HUMAN[All Fields] --> Retrieve 1 Document

Number of documents to display per page: 1000 Mod. Date limit: No Limit

Netscape: PubMed protein query
File Edit View Go Communicator Help
Bookmarks Location: http://www.ncbi.nlm.nih.gov/htbin-post/

3136334
(AF064105) cdc14b2 phosphatase [Homo sapiens]
gi|3136334 [3136334]
(View GenPept Report,FASTA report,ASN.1 report,Graphical view,290 protein neighbors, or 1 nucleotide link)

Netscape: PubMed protein query
File Edit View Go Communicator Help
Bookmarks Location: http://www.ncbi.nlm.nih.gov/htbin-post/

291 citations found

Display structure links for the articles selected (default all).

Netscape: PubMed structure query
File Edit View Go Communicator Help
Bookmarks Location: http://www.ncbi.nlm.nih.gov/htbin-post/

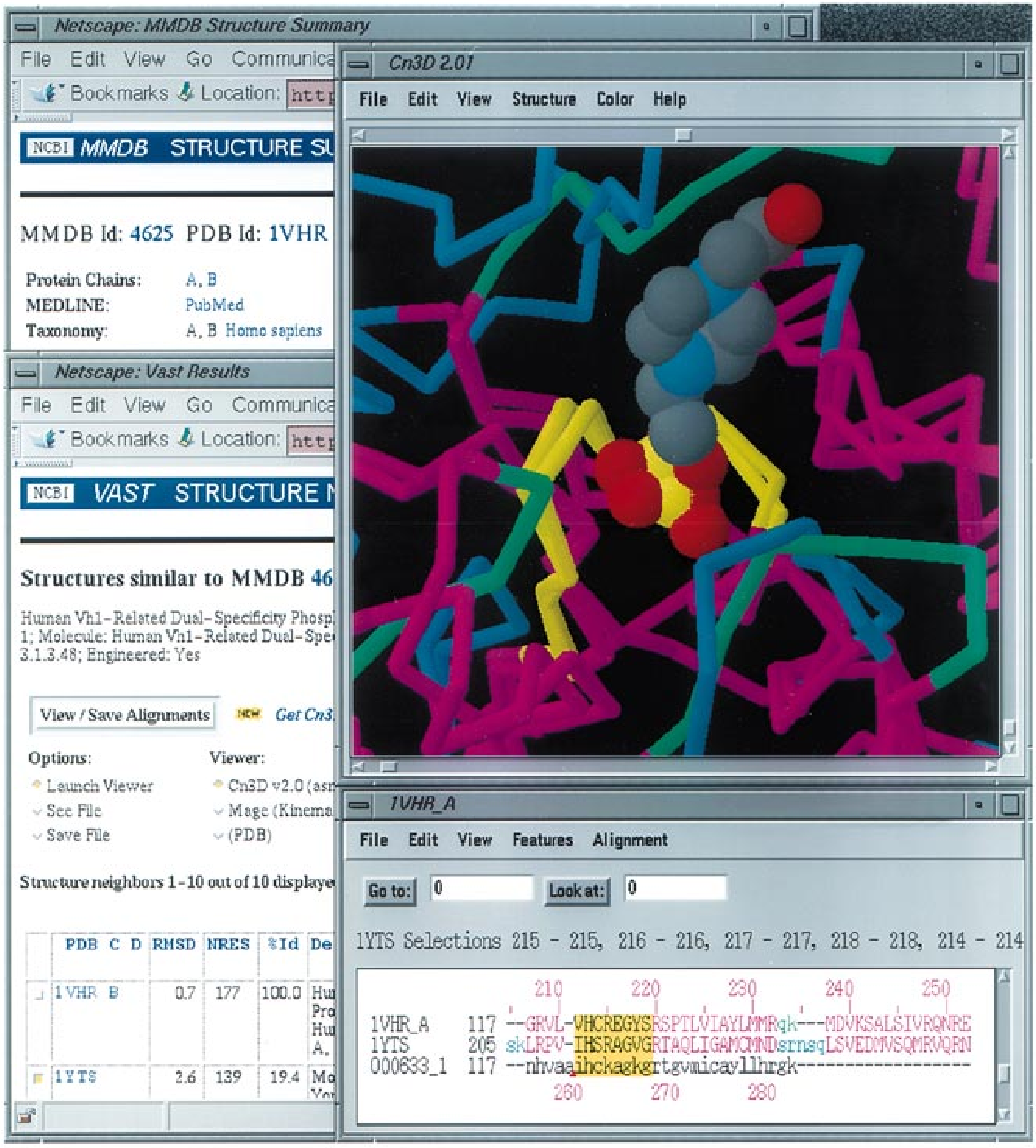
Synonym: D1; Ec: 3.1.3.48; Engineered: Yes; Biological_unit: May Be Inactivated By Dimerization
[5838]
(View Structure Summary,1 MEDLINE link,2 protein links, or 15 structure neighbors)

1VHR
Human Vh1-Related Dual-Specificity Phosphatase Hydrolase, Protein Dual-Specificity Phosphatase Mol_id: 1; Molecule: Human Vh1-Related Dual-Specificity Phosphatase Vhr; Chain: A, B; Synonym: Vhr; Ec: 3.1.3.48; Engineered: Yes
[4625]
(View Structure Summary,3 MEDLINE links, 2 protein links, or 36 structure neighbors)

Figure 1. Screen snapshots summarizing the search for PTEN homologs with 3D structure. Protein Neighbors of PTEN_HUMAN are retrieved from Entrez, with the number of Documents per Page set rather high. From among the neighbors of PTEN_HUMAN, Cdc14b2 is selected, and its sequence neighbors are again displayed. Requesting ’Structure Links’ for any of these 291 sequences identifies the 3D structure 1VHR, a phosphatase homologous to the PTEN gene product.

To examine the 1VHR 3D structure one links to Entrez’s Structure Summary and then chooses ‘View’. This opens viewing windows for structure and sequence such that one may highlight selected residues in the 1VHR sequence, to see where they fall in the 1VHR 3D structure, and vice versa. Viewing windows are managed by Cn3D (8), an Entrez helper application one may download from hotlinks on the Structure Summary page. Examination of 1VHR in this way immediately suggests the location of the phosphatase active site, since the structure contains a bound substrate analog. Pubmed links for 1VHR confirm this interpretation: an article describing mutagenesis results shows that the cysteine residue in the phosphatase site is essential for activity and involved in formation of a phosphocysteine intermediate (9).

To define or better characterize function-associated sites in a 3D structure it is also useful to examine their evolutionary conservation. For this purpose Entrez provides a structure neighbor database, results of pre-computed VAST (10) searches comparing 3D coordinates of each structure to every other. Choosing the Structure Neighbors of 1VHR chain *A*, one sees that there are a number of structures similar to 1VHR in Entrez’s 3D database, including 1YTS (11), a rather distantly related phosphatase showing only 19% sequence identity in VAST’s structure–structure alignment. Selecting 1YTS and View, one may see that



MMDB Id: 4625 PDB Id: 1VHR

Protein Chains:	A, B
MEDLINE:	PubMed
Taxonomy:	A, B Homo sapiens

Structures similar to MMDB 46

Human Vh1-Related Dual-Specificity Phosp
1; Molecule: Human Vh1-Related Dual-Spe
3.1.3.48; Engineered: Yes

Options:	Viewer:
● Launch Viewer	● Cn3D v2.0 (asn
○ See File	○ Mage (Kinema
○ Save File	○ (PDB)

Structure neighbors 1-10 out of 10 displaye

	PDB C D	RMSD	NRES	%Id	De
□	1VHR B	0.7	177	100.0	Hu Pro Hu A,
■	1YTS	2.6	139	19.4	Mo Yer

Go to: 0 Look at: 0

1YTS Selections 215 - 215, 216 - 216, 217 - 217, 218 - 218, 214 - 214

```
1VHR_A    117 --GRVL-VHCREGYSRSPTLVIAYLMMRqk---MDVKSALSIVRQNRE
1YTS      205 skLRPV-IHSRAGVGRIAQLIGAMCMNDsrnsqLSVEDMVSQMRVQRN
000633_1  117 --nhvaaihckagkgrtgvmicayllhrgk-------------------
```

Figure 2. Screen snapshots summarizing structure-neighbor alignments for 1VHR and 1YTS and alignment of the PTEN sequence with 1VHR. The Cn3D display of the backbone trace is colored magenta in regions where VAST has detected similarity between the 3D structures of 1VHR and 1YTS. This includes the apparent phosphatase active site. To check whether residues in this region are conserved in PTEN the sequence PTEN_HUMAN has been imported into this alignment. Conserved residues around the phosphatase active site have been highlighted yellow in both sequence and structure windows. One may see that the active-site cysteine in 1VHR is conserved in PTEN.

the structure (and sequence) surrounding the apparent phosphatase site in 1VHR is well conserved in 1YTS, and that a bound sulfate ion in 1YTS occupies nearly the same site as the bound substrate analog in 1VHR. One may conclude that the fine structure of the apparent active site is conserved in this phosphatase family.

VIEWING SEQUENCE–STRUCTURE ALIGNMENTS

To ask whether sequence features associated with phosphatase activity in 1VHR (and 1YTS) are conserved in PTEN one must examine the PTEN/1VHR alignment in detail. To support this analysis Entrez's visualization system allows one to 'import' a new sequence into the viewing window, aligning it with sequences already displayed there. To align the PTEN gene product sequence with the sequence of 1VHR one chooses (in the sequence viewing window) <File><Download from Entrez><Blast>, and then types 'PTEN_HUMAN'. Importing the PTEN sequence into the alignment of 1VHR and 1YTS, one sees that there is local sequence conservation around the apparent phosphatase site. As may be seen in Figure 2, the active-site cysteine of 1VHR, in particular, appears to be conserved in the PTEN gene product.

This analysis suggests that the phosphatase activity of the PTEN gene product depends on a mechanism similar to that of

1VHR, and that the fine structure of the PTEN phosphatase site may be essential for this activity. Recalling that point mutations associated with Cowden Disease are known, one may now ask whether there is any correspondence between the disease-associated sites and the apparent phosphatase active site. Referring back to the sequence entry PTEN_HUMAN, one may see that the answer is yes: residues 123, 124 and 129 in PTEN are associated with Cowden Disease, and they are also aligned with conserved active-site residues in the 1VHR phosphatase. The disease-associated mutation C124R in fact corresponds to the C124S mutation shown in the case of 1VHR to abolish phosphatase activity (8).

While this simple analysis cannot explain the molecular pathology of Cowden Disease, it nonetheless suggests an interesting possibility: some of the disease-associated mutations in PTEN may directly affect the gene product's phosphatase activity, and perhaps abolish it. As it turns out, this is exactly the mechanism previously suggested as the basis of PTEN's activity as a tumor suppressor (1,12). We hope this example illustrates how one may use Entrez's 3D structure database to answer questions such as this concerning the structure and function of newly discovered proteins.

OTHER FEATURES AND AVAILABILITY

Entrez's 3D database has other query and visualization features not described here. One may directly query the 3D database according to a number of fielded data items, such as text descriptions, author names, dates, journal names, or the taxonomic assignments of Entrez's 3D structures. Entrez's Cn3D visualization system also supports rendering and complexity settings not described here (8). Certain visualization services are also supported by the RasMol (13) and Mage (14) viewers, which may also be installed as helper-applications for WWW browsers. In searching for homologs with 3D structure and interpreting this information, however, it is Entrez's neighbor databases and alignment visualization system that provide uniquely powerful tools.

Entrez's 3D structure database is currently updated once per month. It contains all structures in the PDB (5), with the exception of theoretical models. Entrez's structure neighbor database is updated on the same schedule, but with a lag of roughly one week, due to the computer time required for structure–structure comparison and alignment. For researchers wishing to perform their own computations both structure and structure–structure alignment data in the ASN.1 format used by Entrez are available for bulk download. Source code for the Cn3D visualization system is also available to download as part of the NCBI toolkit library, should others wish to use parts of this system in other software.

Entrez may be accessed at http://www.ncbi.nlm.nih.gov/Entrez/ . The helper-application Cn3D may be downloaded for PC, Macintosh, and UNIX at http://www.ncbi.nlm.nih.gov/Structure/cn3d.html . Comments, suggestions and questions are welcome and should be addressed to: info@ncbi.nlm.nih.gov.

REFERENCES

1 Liaw,D., Marsh,D.J., Li,J., Dahia,P.L., Wang,S.I., Zheng,Z., Bose,S., Call,K.M., Tsou,H.C., Peacocke,M., Eng,C. and Parsons,R. (1997) *Nature Genet*., **16**, 64–67.
2 Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) *Methods Enzymol*., **266**, 141–162.
3 Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F.F., Rapp,B.A. and Wheeler,D.L. (1998) *Nucleic Acids Res*., **27**, 12–17.
4 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res*., **26**, 38–42.
5 Abola,E.E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) In Allen,F.H., Bergerhoff,G. and Sievers,R. (eds), *Crystallographic Databases—Information Content*, *Software Systems*, *Scientific Applications.* Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, pp. 107–132.
6 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol*., **215**, 403–410.
7 Yuvaniyama,J., Denu,J.M., Dixon,J.E. and Saper,M.A. (1996) *Science*, **272**, 1328–1331.
8 Hogue,C.W. (1997) *Trends Biochem. Sci*., **22**, 314–316.
9 Zhou,G., Denu,J.M. and Dixon,J.E. (1994) *J. Biol. Chem*., **269**, 28084–28090.
10 Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) *Curr. Opin. Struct. Biol*., **6**, 377–385.
11 Schubert,H.L., Fauman,E.B., Stuckey,J.A., Dixon,J.E. and Saper,M.A. (1995) *Protein Sci*., **4**, 1904–1913.
12 Nelen,M.R., van Staveren,W.C., Peeters,E.A., Hassel,M.B., Gorlin,R.J., Hamm,H., Lindboe,C.F., Fryns,J.P., Sijmons,R.H., Woods,D.G., Mariman,E.C., Padberg,G.W. and Kremer,H. (1997) *Hum. Mol. Genet*., **6**, 1383–1387.
13 Sayle,R.A. and Milner-White,E.J. (1995) *Trends Biochem. Sci*., **20**, 374.
14 Richardson,D.C. and Richardson,J.S. (1992) *Protein Sci*., **1**, 3–9.